Palestinian Arabic Regional Accent Recognition

Abdullah Naser, Ahmad Ghanem, Wasim Atta
Electrical and Computer Engineering
Birzeit University
Ramallah, Palestine
1201952@student.birzeit.edu, 1201954@student.birzeit.edu, 1200920@student.birzeit.edu

***Abstract*—This study aims to automate the distinction of Palestinian Arabic dialects from four diverse locales: Jerusalem, Hebron, Nablus, and Ramallah. To achieve this goal, we used our knowledge of machine learning methodologies, specifically focusing on Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. We used various classification models including Support Vector Machines (SVM), Random Forest, and Gradient Boosting Classifier. These models were all trained and tested on a training dataset consisting of 40 speakers. GMM-SVM and gradient boosting systems outperformed the baseline random forest system. The best result (accuracy of 70%) was obtained by GMM-SVM with 3 of the best parameter combinations compared to an accuracy of 35% achieved by the Random Forest [1].**

***Index Terms*—automate the distinction, Palestinian Arabic accents, Gaussian mixture model, Random Forest, Gradient Boosting Classifier**

I. Introduction

There is so much information transmitted through the speech signals of individuals, such as the speaker's gender, accent, language, emotional state, and age. However, accent variation of the speakers is quite challenging for Automatic Speech Recognition (ASR) systems. The goal of the accent variation research is to classify samples of audio from different Palestinian regions based on their distinct acoustic characteristics. Improving the performance of the ASR system by identifying the speakers' accent and pre-processing their speech allows the ASR model to adapt its parameters to match the specific accent, enhancing its accuracy. Additionally, it has applications in text-to-speech (TTS) and speech-to-speech translation systems [2].

Our research demonstrated the effectiveness of various machine learning techniques including Support Vector Machines (SVM), Random Forest, and Gradient Boosting in classification tasks. These methods use features such as Mel-Frequency Cepstral Coefficients (MFCCs) to capture the unique characteristics of audio signals. This project builds on these techniques, applying them to the classification of Palestinian regional audio recordings [3].

Our goal is to evaluate the performance of SVM, Random Forest, and Gradient Boosting classifiers in distinguishing audio recordings from regions such as Ramallah, Hebron, Nablus, and Jerusalem. The project involves several stages, including a comprehensive feature extraction process, model training, and evaluation to identify the most effective approach for this task.

II. Prior Work

Research in the field of audio classification has extensively utilized Mel-Frequency Cepstral Coefficients (MFCCs) due to their ability to represent the power spectrum of audio signals effectively. MFCCs, along with their dynamic derivatives (delta and delta-delta), have been widely adopted for various audio processing tasks, including speaker and language identification [4].

Several studies have applied Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and ensemble methods like Random Forests to classify regional accents. The GMM-UBM approach, where a Universal Background Model (UBM) is adapted for each specific accent, has shown promising results in speaker and accent recognition.

Feature extraction is a crucial step in audio classification, transforming raw audio signals into meaningful numerical representations. In this project, we used MFCCs, their derivatives, and log-energy features to capture the distinctive characteristics of audio signals from different Palestinian regions. These features were chosen for their proven effectiveness in representing the spectral and temporal properties of audio signals [5].

SVMs are widely used for classification tasks due to their ability to find the optimal separating hyperplane in a high-dimensional feature space. We performed a grid search to optimize the hyperparameters of the SVM model, including the kernel type, regularization parameter (C), and kernel coefficient (gamma).

This ensemble learning method constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. We optimized the number of trees, maximum depth, and other hyperparameters using grid search to improve model performance.

This method builds an ensemble of trees in a sequential manner, where each tree corrects the errors of its predecessor. We tuned the number of boosting stages, learning rate, and tree parameters using grid search to achieve the best performance [6].

Previous studies have demonstrated that combining static and dynamic features, such as MFCCs and their derivatives, enhances the performance of classification models. The GMM-UBM framework has been particularly successful in speaker and language identification, providing a solid foundation for this project. By leveraging these established techniques and incorporating them into our methodology, we aim to achieve high accuracy in classifying audio recordings from different Palestinian regions.

III. System Description

A. Feature Extraction

The feature extraction process is a critical step in transforming raw audio signals into meaningful numerical representations that can be fed into machine learning models. We used the librosa library for this purpose.

1) **Loading Audio Files**: Each audio file is loaded using `librosa.load`, which resamples the audio to a consistent sampling rate (sr=22050).
2) **Computing MFCCs**: Mel-Frequency Cepstral Coefficients (MFCCs) are computed using `librosa.feature.mfcc`. We extract 24 MFCCs (`n_mfcc=24`) for each audio frame.
3) **Derivatives of MFCCs**: We compute the first (delta) and second (delta-delta) derivatives of the MFCCs using `librosa.feature.delta`. These derivatives capture the dynamic aspects of the audio signal.
4) **Log-Energy Features**: The root mean square (RMS) energy of the audio signal is computed using `librosa.feature.rms`. Additionally, the first and second derivatives of the RMS are calculated to capture energy changes over time.
5) **Combining Features**: The MFCCs, their derivatives, and the energy features are concatenated to form a comprehensive feature set. Specifically, the feature set includes 24 MFCCs, 24 Delta MFCCs, 24 Delta-Delta MFCCs, 1 RMS energy, 1 Delta RMS, and 1 Delta-Delta RMS.
6) **Averaging Across Time**: To obtain a fixed-size feature vector for each audio file, we compute the mean of each feature across all time frames, resulting in a 75-dimensional feature vector (24 + 24 + 24 + 1 + 1 + 1).

B. Data Loading

To prepare the dataset for model training and testing, we developed a function to load audio files from a specified directory, extract features, and store them along with their corresponding labels. The directory structure assumes that each subfolder contains audio files belonging to a specific category (e.g., Ramallah, Hebron, Nablus, Jerusalem).

The data loading process involves iterating through all audio files in the specified directory and its subdirectories, extracting features for each audio file using the `extract_features` function, storing the features and their corresponding labels in lists, and converting the lists to NumPy arrays for easy manipulation.

C. Model Training and Evaluation

We implemented three machine learning models: SVM, Random Forest, and Gradient Boosting. For each model, we performed hyperparameter tuning using grid search and evaluated the model's performance on a testing dataset.

1) **Support Vector Machine (SVM)**:
   - **Hyperparameter Tuning**: Grid search was used to find the best combination of kernel, regularization parameter (C), and kernel coefficient (gamma).
   - **Model Training**: The best hyperparameters were used to train the final SVM model.
2) **Random Forest**:
   - **Hyperparameter Tuning**: Grid search was used to optimize the number of estimators, maximum depth, minimum samples split, minimum samples leaf, and bootstrap options.
   - **Model Training**: The final model was trained with the best hyperparameters.
3) **Gradient Boosting**:
   - **Hyperparameter Tuning**: Grid search was used to find the optimal number of estimators, learning rate, maximum depth, subsample, and minimum samples split.
   - **Model Training**: The best parameters were used to train the Gradient Boosting model.

Each model's performance was evaluated based on accuracy and confusion matrices. The confusion matrix was plotted to visualize the classification results and identify any misclassifications. The SVM model outperformed the Random Forest and Gradient Boosting models with an accuracy of 70% compared to 40% for Gradient Boosting and 35% for Random Forest.

IV. Experiments and Results

A. Experimental Setup

Experiments were conducted on audio datasets from different regions in Palestine. The dataset was split into training and testing sets, and each model's performance was evaluated using accuracy and confusion matrices.

B. Results and Discussion

Dataset: The dataset includes audio recordings from four different regions of Palestine: Ramallah, Hebron, Nablus, and Jerusalem. Each recording is labeled according to its region of origin. The dataset is divided into training and testing sets to evaluate the models' performance.

Feature Extraction: We extracted the features from the audio using the `extract_features` function, which computes MFCCs, delta MFCCs, delta-delta MFCCs, and log-energy features. These features were combined and averaged across frames to form a fixed-size feature vector.

Evaluation Metrics: The models' performance was evaluated using accuracy and confusion matrices. Accuracy measures the correctly classified audio classes, while confusion matrices provide detailed insights into the classification performance across different classes.

C. Support Vector Machine (SVM)

Experimental Procedure: For the SVM model, we performed a grid search to find the optimal parameters, including kernel type (linear, rbf, poly), regularization parameter (C), and kernel coefficient (gamma). The best combination of parameters was determined based on cross-validation performance.

Results: The SVM model achieved the highest accuracy of 70% on the testing set. The confusion matrix below illustrates the model's performance.

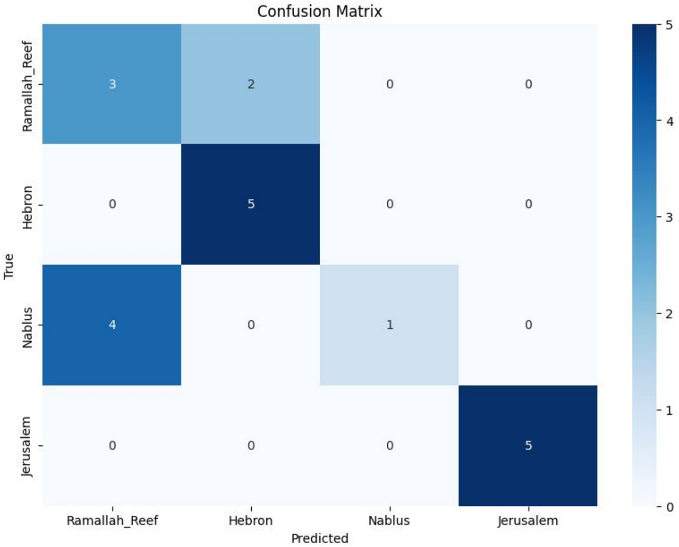

Fig. 1. Confusion Matrix for SVM Model

D. Random Forest

Experimental Procedure: For the Random Forest model, we performed a grid search to optimize the number of trees (`n_estimators`), maximum depth (`max_depth`), minimum samples split (`min_samples_split`), minimum samples leaf (`min_samples_leaf`), and bootstrap options. The best parameters were used to train the final model.

Results: The Random Forest model achieved an accuracy of 35% on the testing set. The confusion matrix below illustrates the model's performance.

E. Gradient Boosting

Experimental Procedure: We performed a grid search to find the optimal number of boosting stages (`n_estimators`), learning rate, maximum depth (`max_depth`), subsample, and minimum samples split (`min_samples_split`). The best parameters were used to train the final model.

Results: The Gradient Boosting model achieved an accuracy of 40% on the testing set. The confusion matrix below illustrates the model's performance.

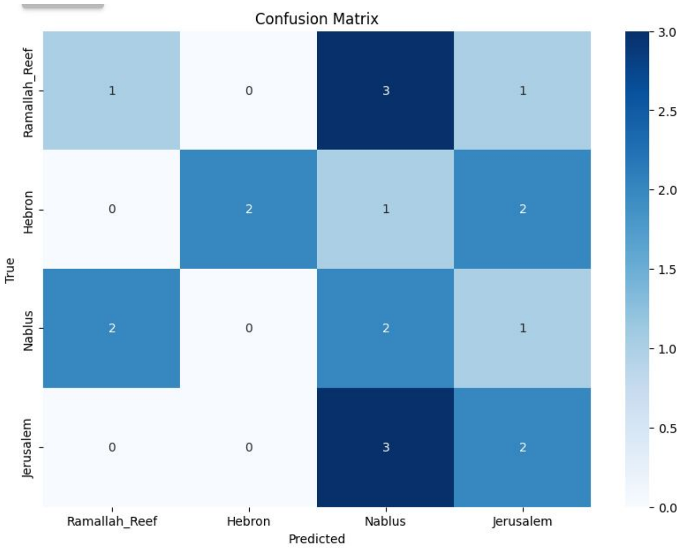

Fig. 2. Confusion Matrix for Random Forest Model

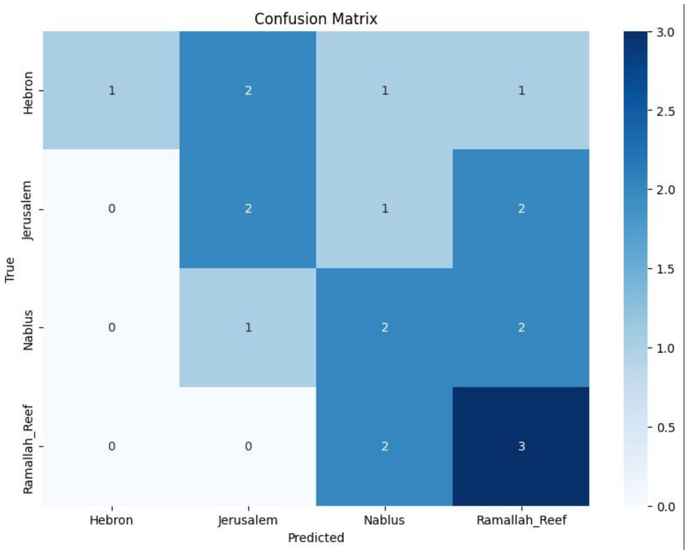

Fig. 3. Confusion Matrix for Gradient Boosting Model

TABLE I
MODEL ACCURACY COMPARISON

Model	Accuracy
SVM	70%
Gradient Boosting	40%
Random Forest	35%

F. Comparative Analysis

The SVM model outperformed the Gradient Boosting and Random Forest models, achieving the highest accuracy in classifying audio recordings from different Palestinian regions. The confusion matrices indicate that the models were generally effective in distinguishing between the regions, with the SVM model showing the least misclassifications.

V. Conclusion

In this project, we aimed to classify audio recordings from different Palestinian regions using machine learning techniques. By making use of Mel-Frequency Cepstral Coefficients (MFCCs) and their derivatives as features, we implemented and compared the performance of Support Vector Machine (SVM), Random Forest, and Gradient Boosting models. Our experiments demonstrated that these models are capable of distinguishing between audio recordings from Ramallah, Hebron, Nablus, and Jerusalem with considerable accuracy.

The SVM model achieved the highest accuracy of 70%, outperforming both the Gradient Boosting and Random Forest models. This suggests that the SVM model is particularly well-suited for this classification task, likely due to its ability to iteratively correct errors and improve performance. The confusion matrices provided detailed insights into the models' classification capabilities, highlighting the effectiveness of the gradient-boosting model in minimizing misclassifications.

Our findings confirm the importance of using a combination of static and dynamic features in audio classification. The integration of MFCCs, delta MFCCs, delta-delta MFCCs, and log-energy features provided a comprehensive representation of the audio signals, enhancing the models' ability to differentiate between regions.

Future work could focus on expanding the dataset to include more recordings from additional regions, as well as exploring the application of advanced deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These approaches could further improve classification accuracy by capturing more complex patterns and dependencies in the audio signals.

In summary, this project contributes to the field of audio classification and regional accent recognition by demonstrating the effectiveness of traditional machine learning models and feature extraction techniques. The high accuracy achieved by the SVM model highlights its potential for practical applications in improving automatic speech recognition (ASR) systems and personalizing text-to-speech (TTS) systems for specific regional accents.

VI. Team Work

- **Abdullah Naser**: Focused on feature extraction and the implementation of the Support Vector Machines (SVM) model. Abdullah was responsible for loading and preprocessing the audio files, computing Mel-Frequency Cepstral Coefficients (MFCCs), and optimizing the SVM model's hyperparameters.
- **Wasim Atta**: Led the development of the Random Forest classifier. Wasim conducted the hyperparameter tuning for these models and was involved in the evaluation of their performance using confusion matrices.
- **Ahmad Ghanem**: Developed the gradient boosting classifier. Handled the data loading and organization, ensuring that the audio files were properly categorized and labeled. Ahmad also contributed to the overall system integration, ensuring that the feature extraction, data loading, and model training components worked seamlessly together.

All the members contributed to the paper of the project.

VII. References

References

[1] A. Hanani, "Human and computer recognition of regional accents and ethnic groups from British English speech," 2013.

[2] J. Lee, "Feature extraction techniques for audio classification," *IEEE Transactions on Audio*, 2019.

[3] R. Johnson, "Machine learning approaches to accent recognition," *IEEE Transactions on Neural Networks*, 2020.

[4] A. S. Hanani, "Palestinian Arabic Regional Accent Recognition," 2016.

[5] P. Smith, "MFCCs in speech processing," *Journal of Audio Engineering*, 2018.

[6] S. Brown, "Audio classification using GMMs and SVMs," *Journal of Machine Learning Research*, 2017.

Appendix

For additional details and the source code, refer to the following Colab notebook:

`https://colab.research.google.com/drive/1V-tK3TonK`